Supplementary Materials for PSDiffusiion

1. Limitations

1.1. Planar composition constraints in volumetric interaction modeling

While our model achieves sophisticated occlusion handling through discrete 2D layer composition, its planar representation inherently limits the modeling of volumetric spatial interactions. The strict layer-wise depth prioritization (e.g., globally enforcing Layer A to occlude Layer B) cannot capture localized mutual occlusion in 3D space, such as intertwined limbs during human embraces. Static layer ordering further fails to adapt to perspective-dependent depth transitions under viewpoint changes, while non-planar intersections with depth-dependent partial visibility remain unresolved due to reliance on discrete alpha compositing. These constraints stem from our use of discretized depth ordering rather than continuous 3D occupancy field learning.

1.2. Uniform sharpness artifacts from 2D-flattened rendering

Despite achieving exceptional per-layer clarity and lighting consistency, the uniform maximum sharpness across all layers prevents emulation of natural depth-of-field effects (e.g., foreground-background focus gradients). This limitation arises from the absence of depth-aware dynamic blurring, compromising visual realism in scenes requiring strong 3D spatial perception. The persistent all-in-focus appearance introduces planar artifacts, particularly noticeable in multi-layer compositions with complex depth relationships, as the discrete compositing mechanism cannot simulate depth-dependent optical blur.

2. More Experiment Details

For layout harmony and interaction plausibility, we use VLM (Qwen2.5-VL-32B-Instruct) to conduct comprehensive scoring. We use 0,1 scoring instead of a percentage-based scoring because we find it better aligned with human judgement. The prompt for Layout Harmony is [Evaluate the layout harmony of the image comprehensively. Consider multiple dimensions including the rationality of the global layout and the scale of each element.] The prompt for Interaction Plausibility is [Evaluate the interaction plausibility of all elements in the image comprehensively. Consider multiple dimensions including physical contacts among entities, coordination of color and style, and consistency of light and shadows.]

3. User Study

The following Table 1 presents the questionnaire form we made for the investigation in the userstudy part.

4. Computation Complexity

The computational complexity of PSDiffusion is $O(n)$, where n is the number of layers. PSDiffusion uses a global branch to provide layout and compositional information to each layer, avoiding costly pairwise calculations. The global branch leverages a pretrained RGB model with global prompts to guide layer synthesis and maintain coherence. The complexity of each component is detailed below.

Layout Arrangement: PSDiffusion ensures plausible layout arrangements with a cross-attention reweighting module. This module extracts relevant cross-attention maps from the global branch to guide each layer's position ($O(n)$ complexity).

Appearance Coherence: PSDiffusion ensures appearance consistency using a partial joint self-attention block, which makes each layer attend only to relevant region of global branch, also $O(n)$. By contrast, full joint attention among all layers and the global branch would incur $O(n^2)$ complexity.

In principle, PSDiffusion can handle a great number of layers generation, but practical limits from the pretrained base RGB model (limitations in generating too many distinct elements following text prompts only precisely) and training data (up to 6 layers) constrain our current setting. We will remain extending layer support in future work.

5. Dataset Visualization

Benefiting from our human-centric curation pipeline and professional-grade multi-stage refinement, our dataset achieves three key superior qualities: high-fidelity alpha mattes, physics-aware inter-layer interactions, and globally harmonized visual aesthetics. Fig. 1, 2 are visualizations of our dataset compositions.

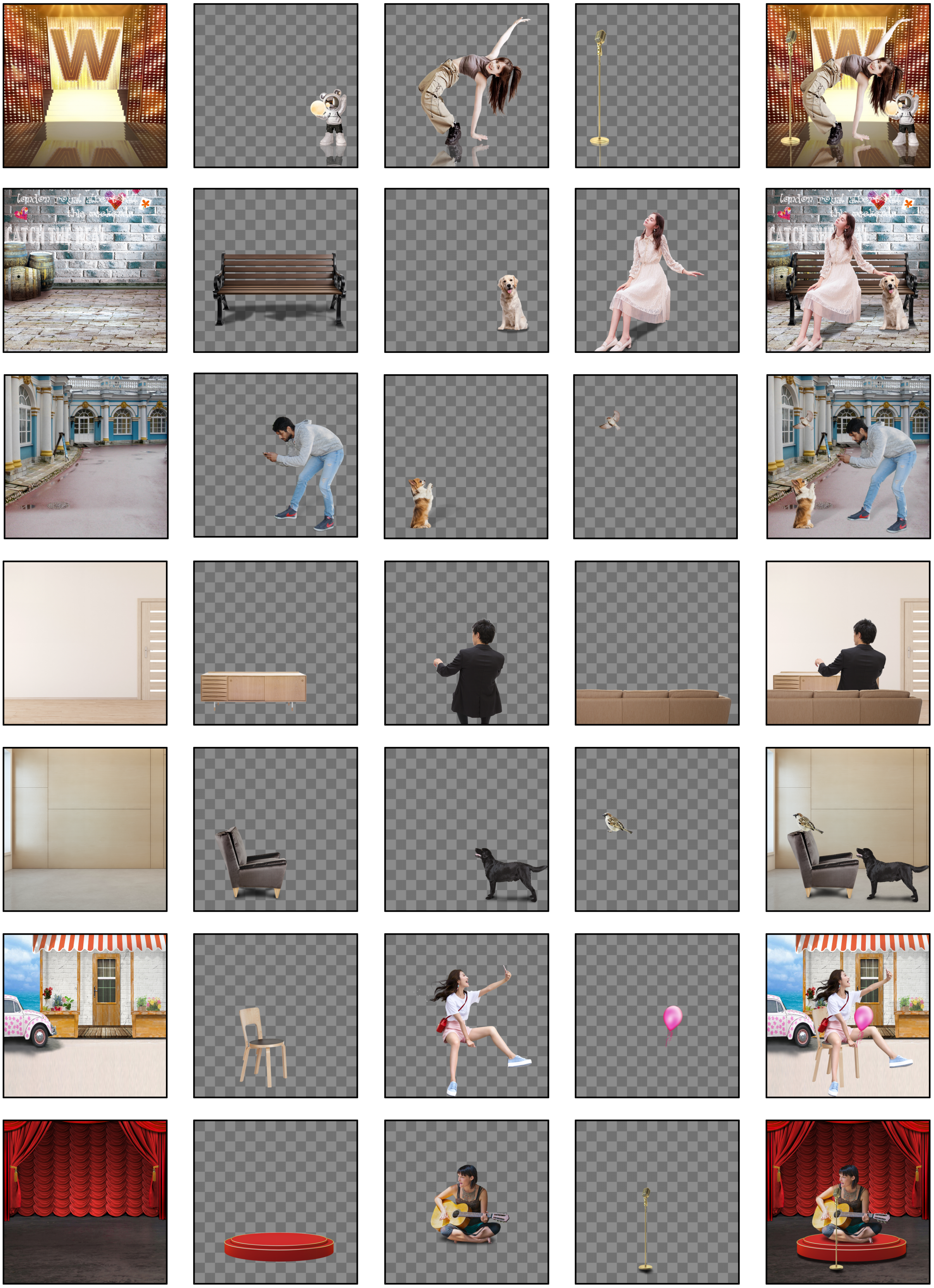


Figure 1. Visualization of our multi-layer dataset

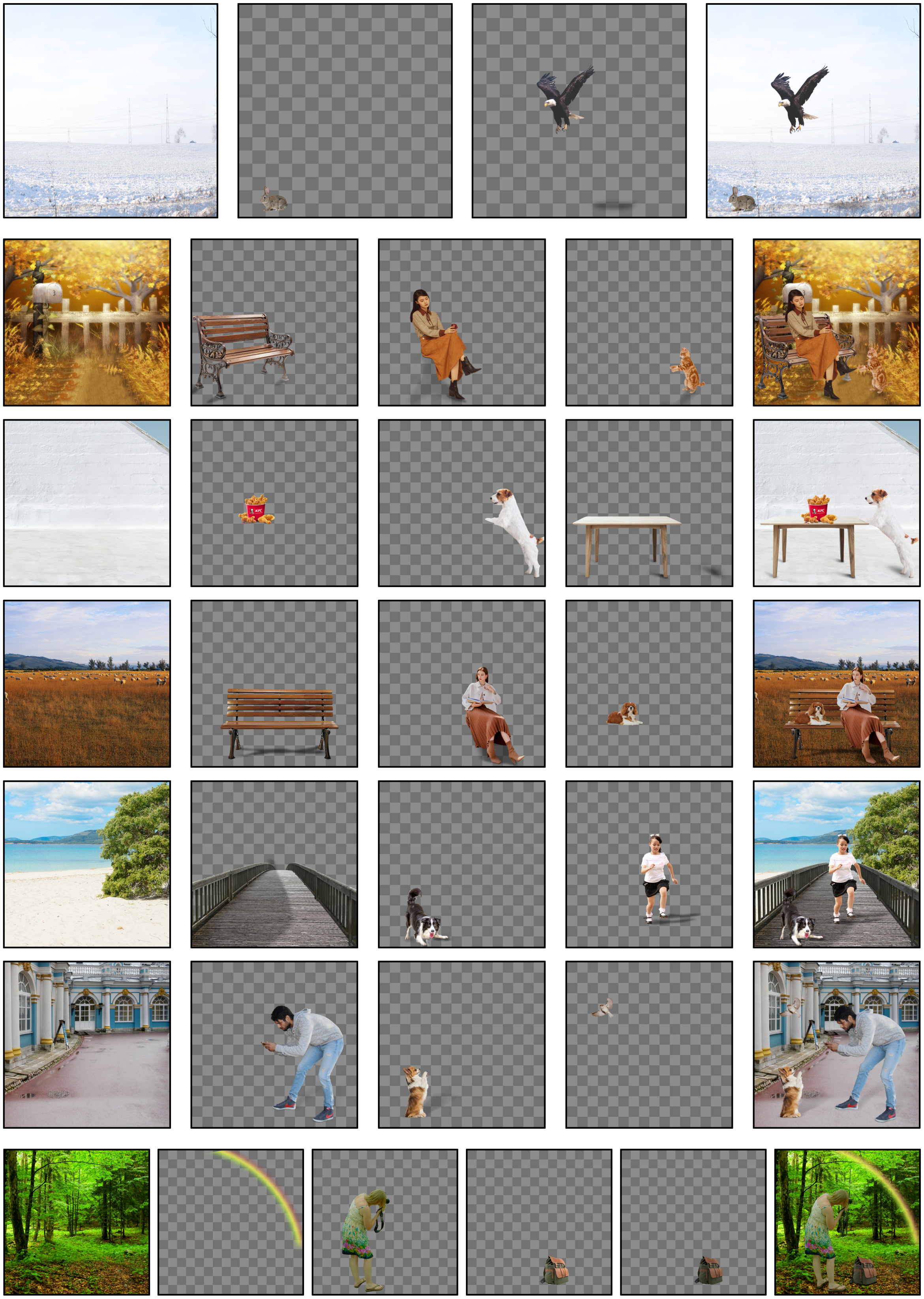


Figure 2. More visualization of our multi-layer dataset

Table 1. Userstudy of PSDiffusion

Current AI image generators create flat images, unlike professional design workflows using layered files (e.g., Photoshop .psd). Our new system automatically generates:

- multi-layer images (1 background + multiple RGBA layers)
- Context-aware layer coordination
- Production-ready layered files

Evaluation Criteria

1. Semantic-textual congruence

- Relevance to prompts & contextual logic
- Visual-perceptual fidelity

2. Aesthetic unity & artifact-free rendering

3. Structural clarity

- Detail preservation & layer definition

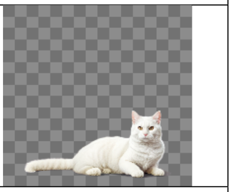
Your Task

For each text-generated image:

- Assign a 1-5 overall score (1=worst, 5=best)
- Base your rating on the three criteria above

Example Case

”whole text”: A white cat sitting and looking at the camera, positioned in front of a wooden floor

Background: wooden floor			
Score			
Foreground: a white cat			
Score			
Merged image			
Score			